

الجمهورية الجزائرية الديمقراطية الشعبية

République Algérienne Démocratique et Populaire

وزارة التعليم العالي والبحث العلمي

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

المدرسة العليا للإعلام الآلي -- 08 ماي 1945 -- بسيدي بلعباس

Ecole Supérieure en Informatique

-08 Mai 1945- Sidi Bel Abbès



Mémoire de Fin d'étude

Pour l'obtention du diplôme d'ingénieur d'état

Filière : Informatique

Spécialité : Intelligence Artificielle et Science des Données (AIDS)

Thème

Mitigating Hallucinations in Large Language Models:
A RAG and Prompt Engineering Approach Applied to Rare Diseases

Présenté par :

- Mr LARBI Mohammed Achraf

Soutenu le : 07/07/2026

Devant le jury composé de :

- | | |
|-------------------------------|--------------|
| • Mme NEMDILI Amel Kahina | Présidente |
| • Mme BELMEKKI Ghizlene Amira | Examinatrice |
| • Mme MEZRAR Samiha | Encadreur |
| • Mme KLOUCHE Badia | Encadreur |
| • Mme Elena Kornyshova | Encadreur |
| • M Olivier Pons | Encadreur |

Année Universitaire : 2025/2026

Abstract

Large Language Models (LLMs) are increasingly applied to clinical decision support, yet their tendency to generate fluent but factually incorrect outputs, termed hallucinations, poses a significant barrier to adoption in safety-critical medical domains. This risk is especially acute for rare diseases, where limited training data and overlapping phenotypes amplify the likelihood of fabricated diagnoses.

This end-of-studies project investigates a hybrid hallucination mitigation approach that combines Retrieval-Augmented Generation (RAG) with structured Prompt Engineering (PE) and knowledge graph enrichment for rare disease diagnosis. A five-source biomedical knowledge corpus (130,316 documents from PubMed, GARD, UMLS, RDS, and PubMed abstracts) was constructed with dual embedding profiles and a three-stage hybrid retrieval pipeline (dense, lexical, and cross-encoder reranking). Twelve distinct diagnostic configurations, spanning prompt-only baselines, RAG variants, graph-augmented architectures, and hybrid CoT-RAG pipelines, were systematically evaluated across two open-source biomedical LLMs (BioMistral-7B and OpenBioLLM-8B) on 200 RareBench cases. A novel Hallucination Gravity Index (HGI) was introduced for severity-weighted assessment.

The best configuration, BioMistral-7B with RAG-Driven Chain-of-Thought, achieved 59.0% exact-match accuracy on a 200-case subset with a full hybrid retrieval pipeline. For reference, Chen et al. [1] report 40.0% for GPT-4 on the full RareBench benchmark without retrieval augmentation; however, a direct comparison is not possible as the two evaluations differ in subset size, matching criteria, and retrieval setup. An ISO 4454 phenopacket ablation, conducted in collaboration with a parallel research theme using real hospitalisation records, demonstrated that clinician-curated phenotypic inputs reduce hallucination severity by 38% in retrieve-first architectures compared to LLM-extracted phenotypes.

Keywords: Large Language Models (LLMs), Hallucination Mitigation, Retrieval-Augmented Generation (RAG), Prompt Engineering, Rare Diseases, Knowledge Graphs, Hallucination Gravity Index (HGI), ISO 4454 Phenopackets.

Résumé

Les grands modèles de langage (LLM) sont de plus en plus appliqués à l'aide à la décision clinique, mais leur tendance à générer des sorties fluides mais factuellement incorrectes, appelées hallucinations, constitue un obstacle majeur à leur adoption dans les domaines médicaux critiques. Ce risque est particulièrement aigu pour les maladies rares, où la rareté des données d'entraînement et le chevauchement phénotypique amplifient la probabilité de diagnostics fabriqués.

Ce projet de fin d'études étudie une approche hybride d'atténuation des hallucinations combinant la Génération Augmentée par la Recherche (RAG) avec l'ingénierie des prompts structurée et l'enrichissement par graphe de connaissances pour le diagnostic des maladies rares. Un corpus biomédical de 130 316 documents provenant de cinq sources (PubMed, GARD, UMLS, RDS et résumés PubMed) a été construit avec des profils d'embedding doubles et un pipeline de recherche hybride à trois étapes. Douze configurations diagnostiques distinctes ont été évaluées systématiquement sur deux LLM biomédicaux open-source (BioMistral-7B et OpenBioLLM-8B) sur 200 cas RareBench. Un nouvel indice de gravité des hallucinations (HGI) a été introduit pour une évaluation pondérée par la sévérité.

La meilleure configuration, BioMistral-7B avec RAG-Driven Chain-of-Thought, a atteint une précision d'exact-match de 59,0% sur un sous-ensemble de 200 cas avec un pipeline de retrieval hybride complet. À titre de référence, Chen et al. [1] rapportent 40,0% pour GPT-4 sur le benchmark RareBench complet sans augmentation par retrieval ; toutefois, une comparaison directe n'est pas possible car les deux évaluations diffèrent en taille de sous-ensemble, critères de correspondance et configuration de retrieval. Une étude d'ablation par phénopackets ISO 4454, réalisée en collaboration avec un thème de recherche parallèle utilisant des dossiers d'hospitalisation réels, a démontré que les entrées phénotypiques curées par des cliniciens réduisent la sévérité des hallucinations de 38% dans les architectures retrieve-first par rapport aux phénotypes extraits par le LLM.

Mots-clés : Grands Modèles de Langage (LLM), Atténuation des Hallucinations, Génération Augmentée par la Recherche (RAG), Prompt Engineering, Maladies Rares, Graphes de Connaissances, Indice de Gravité des Hallucinations (HGI), Phénopackets ISO 4454.